Periodic Signal Extraction with Frequency-Selective Amplitude Modulation and Global Time-Warping for Music Signal Decomposition

Mahdi Triki [#], Dirk T.M. Slock [*], Ahmed Triki [*]

[#] *Philips Research Laboratories, Eindhoven, The Netherlands*
mahdi.triki@philips.com

[*] *Eurecom, Sophia Antipolis, France*
{dirk.slock,ahmed.triki}@eurecom.fr

***Abstract*—A key building block in music transcription and indexing operations is the decomposition of music signals into notes. We model a note signal as a periodic signal with (slow) frequency-selective amplitude modulation and global time warping. Time-varying frequency-selective amplitude modulation allows the various harmonics of the periodic signal to decay at different speeds. Time-warping allows for some limited global frequency modulation. The bandlimited variation of the frequency-selective amplitude modulation and of the global time warping gets expressed through a subsampled representation and parametrization of the corresponding signals. Assuming additive white Gaussian noise, a Maximum Likelihood approach is proposed for the estimation of the model parameters and the optimization is performed in an iterative (cyclic) fashion that leads to a sequence of simple least-squares problems.**

I. Introduction

Sinusoidal model-based music analysis/synthesis has received considerable interest in the computer music community. The sinusoidal transform, originally developed by Quatieri and McAulay [4], represents a signal as a sum of discrete time-varying sinusoids or partials:

$$s(n) = \sum_{k=0}^{P} a_k(n) \cos\left(\theta_k(n)\right) \quad . \tag{1}$$

The estimation of the model parameters is typically carried out using a short-time Fourier transform (STFT) with a £xed analysis frame size and a £xed stride between frames. The sinusoids are extracted by peak-picking in the STFT magnitude spectrum. Intermediate values are obtained by interpolation. A fundamental drawback faced by the traditional sinusoidal-model based techniques, and which arises due to the STFT, is smearing of the frequency response [6], [5]. In fact, over the period of a single analysis frame, the algorithm estimates the amplitude, frequency and phase of any sinusoids it believes to be present. Because of the near logarithmic scale of pitch perception, we need very long windows in order to accurately estimate the pitch of low frequency partials. On the other hand, the time resolution of these parameters is only as £ne as the window length itself. And, since the music signal is highly non-stationary, it is not always possible to £nd a good tradeoff between time and frequency resolution. Also, determining the sinusoid parameters from the STFT peak amplitude and phase only works well for high frequency resolution, high SNR and in the absence of modulation. To overcome the resolution limit of the Fourier transform (due to windowing), non-linear interpolation [7], [9], [8] and dichotomy based approaches [10], [11] were suggested to better localize the peak in the STFT domain. High-Resolution (HR) methods are also proposed to overcome the STFT resolution limit and to provide more accurate estimates of the signal parameters [12].

The major limitation of these techniques is that they ignore the harmonic structure of the music signal. They consider the signal as a mixture of a £nite number of arbitrary sinusoids, and not as a periodic signal. For treating periodic signals, the state of the art is limited to the estimation of pure periodic signals with periodicity equal to an integer number of samples [1], [2]. In these references, the authors propose a Maximum Likelihood approach to analyze pure periodic signals. They show that the resulting procedure can be interpreted as a signal projection onto suitable subspaces. The decomposition of audio signals into periodic features was reconsidered by De Cheveign and Slama [3], and was applied for periodic source separation.

In summary, the drawback of the sinusoidal modeling based techniques is that it considers the signal as a mixture of a £nite number of arbitrary sinusoids (ignoring the harmonic structure of the audio signal); whereas periodic modeling seems to be too rigid to model real audio signals. Motivated by this observation, we have proposed in [13] merging the periodic signal analysis and sinusoidal modeling in order to give more ¤exibility to the periodic signal analysis and impose more structure on sinusoidal modeling. We have considered periodic signals with non-integer period and global amplitude variation and time warping. The use of this model gives a compromise between reality and a parsimonious parametrization. Indeed, global amplitude variation re¤ects mostly attack, sustain, and decay of the whole note signal, whereas global time-warping allows the capture of vibrato and sliding notes. Experimental results reveal that the proposed approach allows extracting

several musical notes accurately from an underdetermined mixture, and produces good auditive synthetic results [14]. Simulations also show that the proposed scheme outperforms the classic separation schemes (based on sparse representation) in terms of accuracy and robustness [15].

A major limitation of the proposed model is that it allows for no spectral variation throughout the note duration, but only amplitude and (synchronized) frequency modulation. Such a model assumes that at any time instant the instantaneous amplitudes and frequencies of the various harmonics of the periodic waveform are proportional. The problem with such a model though is that, in reality, periodic signals produced by musical instruments (e.g. string instruments) have harmonic components that decay at different speeds. Typically higher harmonics decay faster than lower harmonics. In this paper, we introduce a frequency-selective attenuation to alleviate this side effect, and this in a time-varying fashion to re¤ect the time-varying amplitude.

This paper is organized as follows. In sections II and III, the global modulation models (with ¤at and frequency-selective amplitude modulation) and the associated audio signal extraction procedures are presented. Experimental model validation is performed in section IV. Finally, a discussion and concluding remarks are provided in section V.

Notations: upper- and lower-case boldface letters denote matrices and vectors, respectively. As the quantities considered herein are real, $(.)^H$ represents either the transpose, and the complex-conjugate (Hermitian) transpose operators. The symbol T is reserved to denote the assumed period of the audio signal.

II. Audio Modeling with Global Amplitude Modulation and Global Time-Warping

In the sinusoidal modeling, the signal is modeled as a sum of evolving sinusoids as in (1), where $\psi_k(n)$ represents the instantaneous phase of the k^{th} partial. Since the audio signal is almost harmonic, $\psi_k(n)$ can be decomposed into

$$\psi_k(n) = 2\pi k n f_0 \ + \ 2\pi\varphi_k(n) \tag{2}$$

where $\varphi_k(n)$ characterizes the evolution of the instantaneous phases around the k^{th} harmonic, and can be assumed to be slowly time varying. The global modulation assumption implies that all harmonic amplitudes evolve proportionately in time, and that the instantaneous frequency of each harmonic is proportional to the harmonic index, i.e.,

$$\left\{ \begin{array}{l} a_k(n) = a_k \ a(n) \\ 2\pi\varphi_k(n) = 2\pi k \ \varphi(n) \ + \ \Phi_k \end{array} \right. . \tag{3}$$

In summary, we model an audio signal as the superposition of harmonic components with a global amplitude modulation and global time-warping:

$$\begin{array}{l} y(n) \ = \ s(n) \ + \ v(n) \\ = \ a(n) \sum_k a_k \cos\left(2\pi k f_0 \left(n + \frac{\varphi(n)}{f_0}\right) + \Phi_k\right) + v(n) \\ = \ a(n) \ \theta\left(n + \frac{\varphi(n)}{f_0}\right) \ + \ v(n) \end{array} \tag{4}$$

where : - $v(n)$ is additive white Gaussian noise.

- $a(n)$ represents a ¤at amplitude modulating signal.
- $\varphi(n)$ denotes a phase modulating signal (that can be interpreted in terms of time-warping).
- $\theta(n) \ = \ \sum_k a_k \cos\left(2\pi k f_0 n + \Phi_k\right)$ is a periodic signal with a period $T = \frac{1}{f_0}$ (normalized waveshape).

Thus, the audio signal is modeled as a periodic signal with global amplitude and phase modulation. The periodic signal $\theta(n)$ (the normalized waveshape) characterizes the spectral envelope of the audio source. It can be considered as a signature for instrument classi£cation and recognition applications, whereas the amplitude and phase modulating signals ($a(n)$ and $\varphi(n)$) represent respectively the time evolution of the note power and pitch. Remark also that the global phase modulation can be interpreted in terms of dynamic time-warping: it "warps" (stretches or compresses in time) the basic periodic signal $\theta(n)$ to £t the received signal $s(n)$.

In [13], we have expressed the time-warping in terms of an interpolation operation over a basic periodic signal. In sum, the audio signal can be written as:

$$\mathbf{y} = \mathbf{A} \ \mathbf{F}\boldsymbol{\theta} \ + \ \mathbf{v} = \mathbf{s} \ + \ \mathbf{v} \tag{5}$$

where :

- $\mathbf{y} = [y(1) \cdots y(N)]^H$, represents the observation vector.
- $\mathbf{s} = [s(1) \cdots s(N)]^H$, represents the signal of interest.
- $\mathbf{v} = [v(1) \cdots v(N)]^H$, denotes the noise vector.
- $\boldsymbol{\theta} \ = \ [\theta(1) \cdots \theta(\lceil T \rceil)]^H$, characterizes the normalized waveshape over essentially one period
- $\mathbf{A} = diag[a(1) \cdots a(N)]$, is a diagonal matrix representing the global amplitude modulating signal. The global amplitude modulating signal is assumed to be lowpass. Then, $a(n)$ can be down-sampled. The remaining samples can be reconstructed using linear interpolation [14].
- $\mathbf{F}$ is an $N \times \lceil T \rceil$ interpolation matrix characterizing the time-warping. See [13] for a detailed description.

Audio enhancement is performed by adjusting the degrees of freedom (in $\mathbf{A}$, $\mathbf{F}$, and $\boldsymbol{\theta}$) such that the received signal matches the best with the assumed model (in the least-squares sense). The degrees of freedom are estimated in a cyclic fashion. The proposed technique was shown to be effective for musical signal enhancement and separation [15]. Furthermore, the different parameters are related to the three basic features in music sounds: pitch ($\varphi(n)$), intensity ($a(n)$), and timbre ($\theta(n)$). The proposed enhancement technique can also be interpreted as a sum of a scaled, translated and modulated harmonic atom ($\theta(n)$). However, contrary to the classic atomic decomposition approaches, the dictionary is not £xed: the atoms are adapted taking into consideration the structure of the received signal [15].

III. Audio Modeling with Global Frequency-Selective Modulation and Global Time-Warping

In the previous section, we have presented the quasi-periodic signal models with global (¤at) amplitude and frequency mod-

ulation. Such a model allows for no spectral variation throughout the note duration, only for amplitude and (synchronized) frequency modulation. The global amplitude modulation implies that all harmonic amplitudes evolve proportionally in time; whereas the global time-warping emphasizes the signal harmonicity. However, the ratio of the different harmonics (modeled through the basic waveshape $\boldsymbol{\theta}$) is assumed to be constant throughout the whole note duration.

The problem with such a model though is that in reality, periodic signals produced by musical instruments, e.g. string instruments, have harmonic components that decay at different speeds. Typically higher harmonics decay faster than lower harmonics. This means that the global amplitude modulation assumption is not satis£ed.

The assumptions of global amplitude and frequency modulation were introduced to have a parsimonious signal representation. Indeed, the higher the number of parameters per second describing the signal, the noisier the parameter estimates, and consequently the reconstructed signal estimate. Introducing an amplitude modulating signal per harmonic would allow signi£cant degrees of freedom in describing the signal, but would lead to a high parameter rate (the average number of parameters that appear in the description of one second of the signal). An intermediate parameter rate can be obtained by £ltering the periodic signal with a short FIR £lter that can introduce frequency-selective attenuation, and this in a time-varying fashion to re¤ect the time-varying amplitude.

In summary, we model the audio signal as a superposition of harmonic components with global frequency-selective amplitude modulation and global time-warping, i.e.,

$$y(n) = a_n(q)\ \theta\left(n + \frac{\varphi(n)}{f_0}\right)\ +\ v(n) \tag{6}$$

where $a_n(q) = a_{n,L}q^L + \cdots + a_{n,0} + \cdots + a_{n,L}q^{-L}$ is a symmetric zero-phase FIR £lter, $2L+1$ is the amplitude modulating £lter length, and q^{-1} is the time delay operator.

Using matrix notations, the audio signal gets expressed as in (5), where the diagonal matrix $\mathbf{A}$ (characterizing the global amplitude modulation) is replaced by an $L+1$ symmetric band matrix.

The rows of $\mathbf{A}$ contain the coef£cients of the FIR modulating amplitude ($a_n(q)$). This time-varying £lter models the evolution of the note power as well as the relative decay of the different harmonics. Typically, as high frequencies decay faster than low frequencies, the modulating £lter becomes more and more low-pass.

A. Periodic signal extraction procedure

The previous model is linear in $\boldsymbol{\theta}$, $\mathbf{A}$, or $\mathbf{F}$ (separately), $\mathbf{F}$ being parameterized nonlinearly. Trying to estimate all factors jointly is a dif£cult nonlinear problem. Indeed, as the noise is assumed to be a white Gaussian signal, the ML approach leads to the following least-squares problem:

$$\min_{A,F,\theta} \|\mathbf{y} - \mathbf{A}\,\mathbf{F}\,\boldsymbol{\theta}\|^2 \tag{7}$$

where $\mathbf{A}$ and $\mathbf{F}$ are parameterized in terms of subsamples. However, the estimation can easily be performed iteratively, iterating over the following three steps.

1) Periodic signature estimation:

If we assume that the matrices $\widehat{\mathbf{A}}, \widehat{\mathbf{F}}$ are given, the periodic signature $\boldsymbol{\theta}$ can be isolated as

$$\mathbf{y}\ = \widehat{\mathbf{A}}\,\widehat{\mathbf{F}}\ \boldsymbol{\theta} + \mathbf{v} = \mathbf{G}_\theta\ \boldsymbol{\theta} + \mathbf{v}. \tag{8}$$

Then minimizing (7) w.r.t. θ leads to

$$\widehat{\boldsymbol{\theta}} = \left(\mathbf{G}_\theta^H \mathbf{G}_\theta\right)^{-1} \mathbf{G}_\theta^H \mathbf{y}\ . \tag{9}$$

Hence the periodic signature gets estimated using data over the whole note duration.

2) Instantaneous frequency estimation:

The instantaneous frequency and amplitude modulating signals are estimated on a frame-by-frame basis. The length of these time frames T_f and T_a can differ (T_f is typically longer since the frequency varies more slowly than the amplitude). In each frame, the instantaneous frequency is optimized using (10):

$$\begin{cases} \min\limits_{f} \left\|\mathbf{y} - \widehat{\mathbf{A}}\widehat{\mathbf{F}}(f)\widehat{\boldsymbol{\theta}}\right\|^2 \\ \frac{\Delta f}{f_0} \leq \alpha_{max} \end{cases} \tag{10}$$

where Δf denotes the maximum relative frequency variation in the current frame compared to the previous frame, re¤ecting an assumed limited frequency variation rate. The optimal instantaneous frequency value for the current frame is determined from a £nite set of discrete values within this limited range. Simulations show that the minimization problem is locally convex. Thus, the optimization can be performed using the golden section algorithm.

More accurate techniques can be proposed for the instantaneous frequency estimation (High Resolution (HR) methods [12]) or tracking (such as frequency-locked loop signal tracking [16]). However, for enhancement purposes, the gain resulting from this extra processing is very limited as the frequency selective amplitude modulation can compensate for the inaccuracy in the estimation of the instantaneous frequency.

3) Instantaneous frequency-selective amplitude modulation estimation:

If we assume that the normalized waveshape $\theta(n)$ and the time-warping function $\varphi(n)$ (via $\mathbf{F}(f)$) are given, the received signal $y(n)$ is linear with respect to the amplitude modulating £lter coef£cients, i.e.,

$$\begin{aligned} y(n) &= a_{n,0}\breve{\theta}(n) + \sum_{i=1}^{L} a_{n,i}\left(\breve{\theta}(n-i) + \breve{\theta}(n+i)\right) + v(n) \\ &= \left[\breve{\theta}(n)\cdots\breve{\theta}(n-L)+\breve{\theta}(n+L)\right]\begin{bmatrix} a_{n,0} \\ \vdots \\ a_{n,L} \end{bmatrix} + v(n) \\ &= \qquad\breve{\boldsymbol{\theta}}(n)\qquad\qquad \mathbf{a}(n)\ +\ v(n) \end{aligned}$$

where $\breve{\theta}(n) = \theta\left(n + \frac{\varphi(n)}{f_0}\right)$ is the warped normalized waveshape. Thus, using the current estimates of $(\widehat{\mathbf{F}}, \widehat{\boldsymbol{\theta}})$, the observation vector $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{G}_a \ \mathbf{a} + \mathbf{v}$$

where $\mathbf{G}_a$ is a $N \times (N(L+1))$ block diagonal matrix, and $\mathbf{a} = [\mathbf{a}(1)^H \cdots \mathbf{a}(N)^H]^H$ is a $(N(L+1)) \times 1$ vector characterizing the amplitude modulation.

On the other hand, the coeff£cients of the frequency-selective modulating £lter signals are assumed to be lowpass. Therefore, $\{a_{n,i}\}_{i=0:L}$ can be down-sampled. The remaining samples can be estimated using linear interpolation, i.e.,

$$\mathbf{a}_i = \begin{bmatrix} a_{1,i} \\ a_{2,i} \\ \vdots \\ \vdots \\ a_{N,i} \end{bmatrix} = \begin{bmatrix} 1 & 0 \cdots & 0 \\ P_{21} P_{22} \cdots & 0 \\ P_{31} P_{32} \cdots & 0 \\ 0 & 1 \cdots & 0 \\ & \vdots & \\ 0 & \cdots \cdots & 1 \end{bmatrix} \begin{bmatrix} a_{1,i} \\ a_{T_a+1,i} \\ \vdots \\ a_{N,i} \end{bmatrix} = \mathbf{P}_a \ \mathbf{a}_{\downarrow i}$$

where $\mathbf{a}_{\downarrow i}$ contains the i^{th} coeff£cients of the frequency selective modulating £lter $a_n(q)$, downsampled by the factor T_a. $\mathbf{P}_a$ represents the interpolation matrix used to reconstruct $\mathbf{a}_i$ from its downsampled version $\mathbf{a}_{\downarrow i}$ (see [14] for further discussion on the design of $\mathbf{P}_a$). In summary, the estimation problem can be formalized as:

$$\mathbf{y} = \underbrace{\mathbf{G}_a \left(\mathbf{P}_a \otimes \mathbf{I}_{L+1}\right)}_{\mathbf{G}_{\downarrow a}} \ \mathbf{a}_{\downarrow} + \mathbf{v} \tag{11}$$

where $\otimes$ denotes the Kronecker product, and $\mathbf{a}_{\downarrow} = \left[\mathbf{a}^H(1) \ \mathbf{a}^H(T_a+1) \cdots \mathbf{a}^H(N)\right]^H$ represents the actual degrees of freedom of our model. Thus, the elements of $\widehat{\mathbf{A}}$ are estimated using the least-squares technique (via the estimation of $\mathbf{a}_{\downarrow}$).

IV. Implementation Issues and Experimental Results

We £rst comment on the implementation of the proposed algorithm based on global frequency-selective amplitude and phase modulation (that we refer to as Quasi-Periodic Signal Extraction (QPSE)). Numerical examples are shown next.

The proposed scheme can be implemented in an ef£cient way. In fact by exploiting the sparsity and the structure of the interpolation matrices $\mathbf{F}$ and $\mathbf{P}_a$, we can reduce considerably the required memory and the computation complexity. As we use linear interpolation, each row of the matrices $\mathbf{F}$ and $\mathbf{P}_a$ contains at most two non-zero elements. In addition, for two non-adjacent columns, the sets of the non-zero elements do not overlap. So that, for a $N \times M$ interpolation matrix $\mathbf{P}$ ($\mathbf{F}$ or $\mathbf{P}_a$), the matrix $\mathbf{P}^H\mathbf{P}$ is a tri-diagonal matrix; and the computation complexity of such operation is $4N$ (instead of MN^2). For a given $N \times 1$ vector $\mathbf{y}$, $\mathbf{P}^H\mathbf{y}$ can computed using $2N$ multiplications (instead of MN).

Moreover, one can show that for a given K-band matrix $\mathbf{G}$, $\overline{\mathbf{G}} = \mathbf{P}^H \ \mathbf{G} \ \mathbf{P}$ is a $(K+2)$-band matrix. Thus, to solve the linear system $\overline{\mathbf{G}}\mathbf{x} = \mathbf{b}$, we should consider the LDU decomposition. In such a case, the lower diagonal matrix L in the LDU decomposition is also a banded matrix (£gure 1).

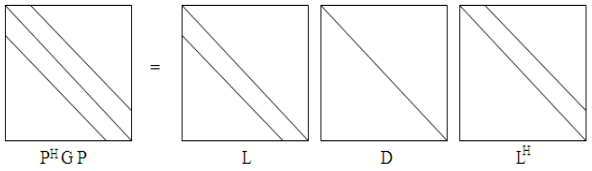

Fig. 1. LDU decomposition of band matix

Once the LDU decomposition is performed, the linear system boils down to sequence of simple forward, instantaneous, and backward triangular systems. Thus, the computational complexity of the solution is $O\left(K^2M\right)$ (instead of $O\left(M^3\right)$).

Next, we validate the proposed extraction approach using real music signals (10 single notes originated from a variety of string and wind instruments). The audio signals were recorded at 44.100 kHz, then downsampled to 16 kHz.

In order to solve the identi£ability problem in (6), we impose that the frequency-selective amplitude modulation is frequency-¤at over a limited portion of the signal, somewhere in the middle. The identi£ability problem arises from the fact that multiplying the amplitude modulating signal by a given £lter $\alpha(q)$ and £ltering $\theta(n)$ by $1/\alpha(q)$ leads to an equivalent decomposition of $y(n)$.

Figure 2 plots the extraction Signal-to-Noise Ratio ($SNR = \frac{\sum_n s(n)^2}{\sum_n (s(n) - \widehat{s}(n))^2}$) using the periodic time-warped model with respectively global frequency-selective amplitude modulation and global ¤at amplitude modulation. The smoothing amplitude modulation factor is set to $T_a = 3T$ ($T = ceil(1/f_0)$ is the period of the harmonic component). No phase modulation was allowed (α_{max} in (10) is set to 0). As expected, the global frequency-selective modulation model £ts better real audio signals and its extraction accuracy increases with £lter length L. In fact, the more coeff£cients the FIR £lter (modeling the frequency-selective amplitude modulation) contains, the more it allows for diverse mode variations, and the better the model £ts the real signal. However, we remarked that for $L \geq 10$ no considerable gain was noticed (10-tap £lter is enough to model the different modes for the tested instruments).

We remark also that the performance of the quasi-periodic signal modeling (with ¤at or frequency-selective amplitude modulation) depends strongly on the harmonicity of the musical instrument. For instance, the model seems not adequate for piano signals. Indeed in stringed instruments such as the piano, the less elastic the strings are (that is, the shorter, thicker, and stiffer they are), the more inharmonicity they exhibit. When a string gets thick enough, compared to its length, it stops behaving as a string and starts acting

more like a cylinder (a tube of mass), which has different harmonics than strings. Moreover, a single piano note attack excites simultaneously 1, 2 or 3 strings (which are in addition not perfectly tuned [1] [20]).

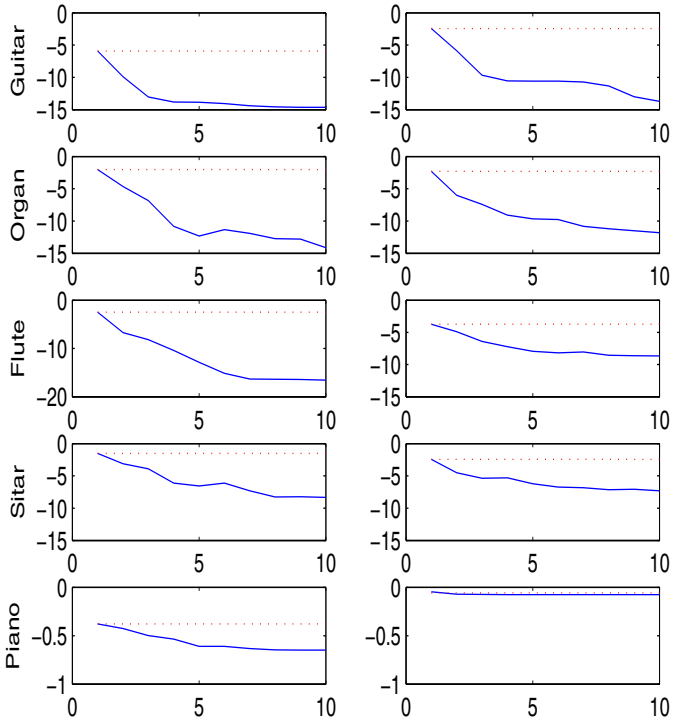

Fig. 2. $-10\log_{10}(\text{SNR})$ vs L, using a periodic time-warped model with global ¤at (dotted line) or frequency-selective (solid line) amplitude modulation .

Next, we investigate the estimation vs. modeling tradeoff. We consider the enhancement accuracy of the proposed scheme in the presence of additive white noise for guitar (£g. 3), ¤ute (£g. 4) and organ (£g. 5) signals.

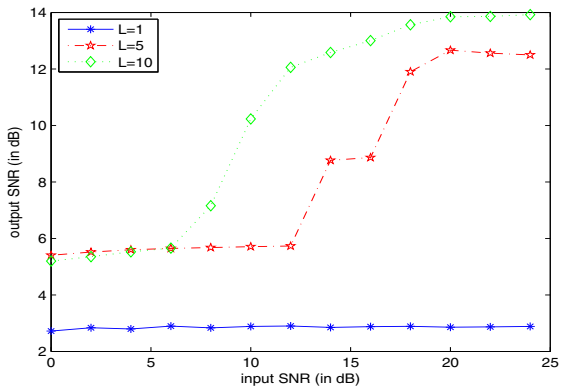

Fig. 3. Guitar signal enhacement SNR using a periodic time-warped model with frequency-selective amplitude modulation ($L = 1,\ 5$ and 10).

Frequency-selective modulation induces additional degrees of freedom. Such a model leads to a parsimonious signal representation (decreasing modeling error). However, the higher the number of the parameters describing the signal, the noisier the parameter estimates and consequently the reconstructed signal estimate. That is why at high SNR, the performances of the frequency-selective modulation increases with L (the estimation error may be neglected). However, at low SNR, $L = 5$ and $L = 10$ produce comparable enhancement accuracy.

[1]Tuning the three strings exactly together gives a tone that not only sounds dead but dies away too rapidly. It also increases the perceived beating in the sound [20].

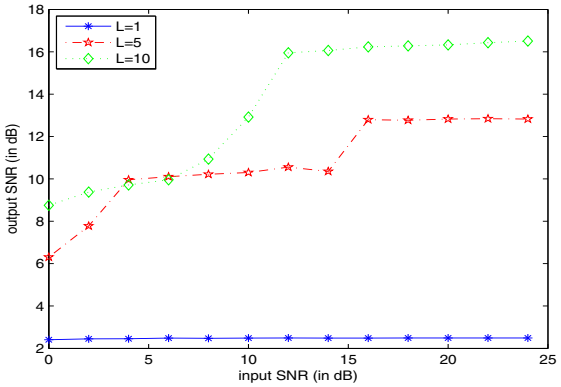

Fig. 4. Flute signal enhacement SNR using a periodic time-warped model with frequency-selective amplitude modulation ($L = 1,\ 5$ and 10).

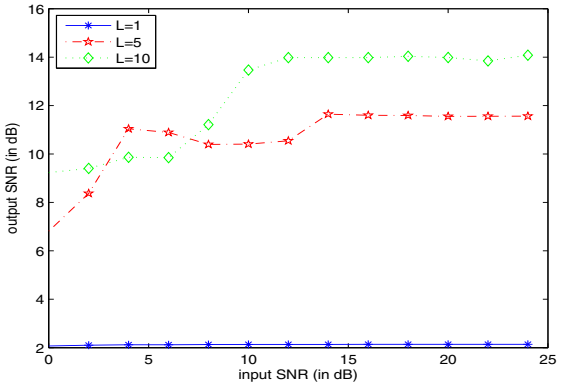

Fig. 5. Organ signal enhacement SNR using a periodic time-warped model with frequency-selective amplitude modulation ($L = 1,\ 5$ and 10).

In sum, simulations show that (for a variety of string and wind instruments) the quasi-periodic signal modeling (with frequency-selective modulation) enables the extraction of the audio signal harmonic component. An interesting application of such approach is music transcription.

Pitch information is an essential part of almost all western music. However, the automatic extraction of the pitch content is a non-trivial problem; and systems trying to perform this task tend to be very complex [17]. Music transcription aims to *detect the 'position'* and to *recognize the 'content'* of the musical event (musical notes and effects such as vibrato, glissando, etc...); which needs both good temporal and frequency resolutions. Comparing to the (frame-by-frame) STFT based approaches [17], [18], the quasi-periodic signal

modeling performs better resolution tradeoff (by exploiting the temporal structure of the musical signal). Indeed, the global amplitude modulation model enables the joint extraction of the different partials, while allowing for slow L decay modes. This fact enhances both note detection and recognition accuracy. In addition, valuable information could be carried out by analyzing the evolution of the amplitude and phase modulating signals (allowing for high temporal resolution transcription, detection of several music effects (vibrato, glissando, etc...).

A key building block in pitch estimation is the evaluation of the salience, or strength, function at the different candidate periods. Classically, the salience is inferred from the spectrum as a weighted sum of the harmonic partials of a given pitch candidate T. Several approaches are proposed to set the weighting coef£cients [17], [18]. In this respect, the extraction SNR (assuming a basic period T) represents an insightful salience measure. Indeed, the QPSE enables the joint extraction of the different partials while imposing a kind of spectral smoothness (over time frames) that has been showed to be valuable (increases the transcription accuracy) [17], [18].

We have tested the monophonic music transcription performance using various instruments (guitar, sitar, ¤ute, and piano). The data (graciously provided by Antony Schutz from Eurecom) was recorded at 44.100 kHz, then downsampled to 22.050 kHz. The maximum number of iterations (in the QPSE cyclic parameters estimation) was £xed to 3. The order of the amplitude modulating £lter was set to $L = 5$. No prior information (about the timbre and/or the instrument) was considered. A standard error metric was used for evaluation [19]: a recall measure (percentage of original notes that were transcribed), and a precision measure (percentage of transcribed notes that were present on the original stream). The average (over all recordings) for each criterion is: recall 98% and precision 100%. Furthermore, 100% of the transcription errors are due to octave mistakes. We remarked also that the transcription accuracy is still quite good even for instruments that present severe inharmonicity (e.g. piano).

V. Concluding Remarks

In this paper, we have investigated signal enhancement techniques exploiting the harmonic structure of the audio signal. We have modeled an audio signal as a periodic signal with (slow) global variation of amplitude (characterizing the evolution of the signal power) and phase (emphasizing the harmonic structure). Time-varying frequency-selective amplitude modulation allows the various harmonics of the periodic signal to decay at different speeds. The bandlimited variation of the frequency-selective amplitude modulation and of the global time warping gets expressed through a subsampled representation and parametrization of the corresponding signals. Simulations show that the extraction technique is suitable for the analysis of several string and wind instruments, and shows good potential for music transcription application.

Acknowledgment

Eurecom research is partially supported by its industrial members: BMW, Bouygues Télécom, Cisco Systems, France Télécom, Hitachi Europe, SFR, Sharp, STMicroelectronics, Swisscom, Thales. The work reported herein was also supported by The European FP6 NoE project K-Space.

References

[1] D.D. Muresan and T.W. Parks, “Orthogonal, Exactly Periodic Supspace Decomposition,” *IEEE Trans. on Signal Processing*, Vol.51, No.9, Sept. 2003.

[2] J.D. Wise, J.R. Caprio, and T.W. Parks, “Maximum Likelihood Pitch Estimation,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 51,pp.418-421, May 1976.

[3] A. de Cheveign and M. Slama, “Acoustic Scene Analysis based on Power Decomposition,” *In Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing* , May 2006.

[4] R. McAulay and T. Quatieri, “Speech Analysis/Synthesis Based on a Sinusoidal Representation,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol.34, Aug. 1986.

[5] P. Prandom, M. Goodwin, and M. Vetterli, “Optimal Time Segmentation for Signal Modeling and Compression,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.3, pp.2029-2032, Apr. 1997.

[6] S.N. Levine, T.S. Verma, and J.O. Smith, “Multiresolution Sinusoidal Modeling for Wideband Audio with Modi£cations,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.6, May 1998.

[7] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, “Minimization or Maximization of Functions,” *Numerical Recipes in C (The Art of Scienti£c Computing)*, chapter 10, pp. 402–405. Cambridge University Press, USA, 1992.

[8] K. Hofbauer, “Estimating frequency and amplitude of sinusoids in harmonic signals-a survey and the use of shifted fourier transforms,” *Master’s thesis, Graz University of Technology*, May 2004.

[9] F. Keiler and S. Marchand,“Survey On Extraction Of Sinusoids In Stationary Sounds,” *In Proc. of Digital Audio Effects Conf. (DAFX)*, pp. 59-64, Sept. 2002.

[10] Y.V. Zakharov and T.C. Tozer,“Frequency Estimator with Dichotomous Search of Periodogram Peak,” *Electronics Letters*, Vol.35, Issue 19, pp.1608-1609, Sep. 1999.

[11] E. Aboutanios,“A modi£ed Dichotomous Search Frequency Estimator,” *Signal Processing Letters*, Vol.11, Issue 2, pp.186-188, Feb. 2004.

[12] R. Badeau, B. David, and G. Richard, “High-Resolution Spectral Analysis of Mixtures of Complex Exponentials Modulated by Polynomials,” *IEEE Trans. on Signal Processing*, Vol. 54, No.4, pp.1341-1350, Apr.2006.

[13] M. Triki and D.T.M. Slock, ‘ ‘Periodic Signal Extraction with Global Amplitude and Phase Modulation for Music Signal Decomposition,” *In Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, Mar. 2005.

[14] M. Triki and D.T.M. Slock, “Multi-channel mono-path periodic signal extraction with global amplitude and phase modulation for music and speech signal analysis,” *In Proc. of IEEE Work. on Statistical Signal Processing (SSP)*, July 2005.

[15] M. Triki and D.T.M. Slock, “Music Source Separation via Sparsi£ed Dictionaries vs. Parametric Models,” *In Proc. of Int. Sym. on Communications, Control, and Signal Processing (ISCCSP)*, March 2006.

[16] A. Wang, “Instantaneous and Frequency-Warped Signal Processing Techniques for Auditory Source Separation,” *Ph.D Thesis*, Stanford University, Aug. 1994.

[17] A. Klapuri, “Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes,” *In Proc. of Int. Symp. on Music Information Retrieval (ISMIR)*, pp.216-221, Oct. 2006.

[18] A. Pertusa, J.M. Inesta,“Multiple Fundamental Frequency Estimation Using Gaussian Smoothness,” *In Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, pp.105-108, Apr. 2008.

[19] G. Reis, N. Fonseca and F. Ferndandez ,“Genetic Algorithm Approach to Polyphonic Music Transcription,” *In Proc. of IEEE Int. Symp. on Intelligent Signal Processing (WISP)*, pp.1-6, Oct. 2007.

[20] A.H. Benade,“Fundamentals of Musical Acoustics,” *Dover Publications*, 1990.